iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41292

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

A Study of Content-Based Recognition of Speech Utterance

Yusra Shafi[1], Satish Saini[2]
[1]*Research Scholar, M.Tech CSE, Department of CSE, RIMT University, Mandi Gobindgarh*
[2]*Professor, Department of ECE, RIMT University, Mandi Gobindgarh, Punjab, India*

Abstract: ***Automated voice recognition is a branch of algorithms and cognitive computing that tries to produce robots which can interact with other folks through speaking. Speaking is a data imply suggesting includes both verbal and psycholinguistic information in addition to paralinguistic data. Emotion is a good example of paralinguistic information transmitted in part through speech. Verbal interaction gets simpler because of the development of technologies that can comprehend paralinguistic information such as mood. In this Cnns circuits' usefulness in language research emotion identification was investigated. Spectrograms with a wide range of frequencies of audio samples remained employed as only a source characteristic for the internet infrastructure The circuits had been educated. to identify speaking patterns provided by performers when expressing a certain mood. We used English-language speech resources to validate and develop our technologies The information used for retraining was expanded by two levels in each database. As a result of this, the dropout strategy was employed to. The gender-agnostic, language-agnostic CNN models, as per our findings, achieved the highest degree of accuracy, beat Consequences of elections in the literary, and equalled or even surpassed human performance on benchmark databases. Future study should test the capability of deep learning methods in voice emotion identification using real-life speech data.***
Keywords: Artificial intelligence, Machine Learning, Speech recognition, Convolutional neural network (CNN)

I. INTRODUCTION

Over the last few decades, numerous research has been done to better explain human brain and to develop artificial intelligence techniques that can mimic human behaviour 1st The adult brain is a complicated organ that has long been the subject of Artificially Intelligent study (AI). Although the human brain can acquire limited perceptual ideas to greater complex notions data, its neural networks are incapable of doing so. The human brain shines in language development, voice comprehension, and face detection when it comes to learning complicated concepts. The primary goal of AI is to develop intelligent machines capable of generating reasonable ideas and activities that are akin to human cognition and behaviour. In the subject of a.i., there are a number of experimental professions known as sub-fields. A few of the major AI fields of research include image processing techniques and machine translation comprehension, as well as pattern recognition, robotics, robot scanning, and pattern recognition. In the future, systems that can understand words will be able to engage with people in a human-like manner. The human brain performs two of the most complicated operations: speaking and understanding speech.

AI research into automatic speech recognition (ASR) aims to create robotics that can converse with mankind via voice [7, 8]. To understand spoken words, early ASR systems relied on linguistic features of speech [9, 10, 11, 12]. One of the early AI jobs was (ASR), sometimes known as voice to text. Speech or spoken language is the most natural way for humans to engage. ASR technology can support interpersonal communication more productive and client. Computational intelligence's actual completion is aided by hardware spoken identification and understanding. A raw human voice is computationally extracted into x1: T of length T.

The most likely decoding hypothesis ω̂, according to Baye's decision rule is given as
This theory can be stated in the form of an M-length sequence.
The decision formula can be stated as follows using Baye's rule:
The acoustic model is denoted by the letters p(x1:T ω |), whereas the language model is denoted by the letters P(). A creative framework is used to describe the ASR system in this arrangement. The acoustic model, p(x1:T |), is estimated using Svm classifier. In the generations following, several significant improvements to the HMM concept have really been made, notably mood tying (Young et al., 1994), racist and prejudicial retraining (Povey and Woodland, 2002), and speaker/noise adjustment (Young et al., 1994). (Povey and Woodland, 2002). (Leggetter and Woodland, 1995; Gale, 1998).

Because of recent improvements in combining deep learning with HMMs, the ASR robots' efficacy has substantially increased Exclusionary models, often known as end-to-end models, have been investigated instead. In these models, neural networks are used to describe the likelihood P(|x1:T), which is directly tied to the decision rule. .. Speech recognition technology has come a long way in the previous half-century. Human habits have begun to change as a result of such technology, and civilization has progressed as a result of it. Nonetheless, there are a number of drawbacks to these approaches. Effective and quick adaptation methods to individual accents and rampant corruption, such as speech affected by noise and reverberation, remain major concerns in current Classification techniques, in especially, are used in Speech recognition systems. In the fields of voice commands and artificial intelligence, they need to be researched further.

II. OBJECTIVES

The major objective of this research is to develop a recognition is a technology a woman's perspective using machine learning techniques and audio formats.

III. LITERATURE REVIEW

Profound learning is a relatively new data mining technique. was created to cope with massive databases and intricate systems. The need for "hand-crafted" criteria prior to categorization has been eliminated thanks to developments in deep learning [13]. These models, also known as Deep Learning (DL), may learn low-level characteristics from training data in their bottom layers before creating high-level models in their top layers. Deep learning models may be utilised since these attributes are automatically extracted. Deep learning algorithms have been increasingly popular in recent years for classifying speech emotions. Several research have investigated at the use of deep learning models in recognising speech mood during the last few years.

Wei-Long Zheng and Bao-Liang Lu employed transferring learning methodologies (TCA based Theme Transferring) to construct Ultrasound efficient ways lacking marked test set that are significantly more exact in identifying pleasant emotions than earlier strategies. Using this strategy, they were able to get this result. Learned was transferred utilising three pillars: TCA, KPCA, and Transudative Attribute Transfer. TCA-based Subject Transfer, KPCA

. They derived DE features from the data after pre-processing A equaliser was used to filter raw EEG data. They used a cross that excluded one participant. -examination when it came time to analyse. Their testing results revealed that the transudative parameter transfer technique surpassed the other alternatives in terms of accuracy, resulting in a 19.58 percent increase in recognition rate. True, this success is limited to the discovery of good candidates. As a result, this method's capacity to detect negative information is restricted. There is still much effort to be done in appropriately distinguishing negative and neutral emotions. [14]

The GAN model has no method of controlling the type of data that will be used. The conditional GAN (CGAN) approach solves this issue by including the ground truth label as an extra parameter in the generator. By making this change, CGAN enables the GAN model to generate new pictures from other classes. The CGAN's module generates an extra class input to decide the new picture type to be created. Only'real' is returned by the determiner when the input seems to be genuine and is aligned to its class supplied by the generator [15].

IV. SYSTEM ANALYSIS

Although deep learning methods (ANNs) are the most well examples of classification methods, several ANNs may be trained using unattended learning techniques. [16]. ANNs are based on biological neural nets, such the central nervous system of humans. To put it another way, they're made up of a lot of networks by adding parts. The basic building blocks of machine learning are on-the-fly networks (ANNs), which are essentially ANNs with many neurons and layers. It's no surprise that deep neural networks are extremely effective in a range of machine learning applications. Deep learning models' capacity to learn complex characteristics from basic data is crucial to their success. The initial layers of deep learning models are built on simple and basic features, often known as basic and simple features. These low-level properties are employed to build a sophisticated representation in the deeper levels. The amount of preparation necessary to extract hand-crafted features prior to building classifiers might be reduced if high-level features are built from low-level features. Deep learning algorithms called convolutional neural networks recognise audio sounds depending on their emotional states.

A. Training algorithm Complexes with Multiple Layers

The structure of the circuit of artificial neural networks is what distinguishes them (ANNs). As a result, it controls the connections between neurons, which are the fundamental processing units The cross-classification model (MLP) was a very well neural network architecture [17] that is built up of input vector units (LTUs).

A linear threshold unit is utilised, as indicated in Figure 1. An LTU accepts weighted inputs (here from three neurons) and calculates, per the example. z = w1x1 +w2x2 + w22+b, to determine the linear set of inputs, use the bias term, b. Following that, a step function such as H(x) or sgn(x) is applied to the linear function to get the result y = f. (z). If the weighted total is larger than a threshold value, the LTU will produce an output (which is determined by the bias term).

On aggregate, an MLP has one input sequence, one or many LTU s (sometimes called cluster centres), and first output layer. Out from encoder (slower pace) to the activation function (wider range), data is transmitted (higher level). As illustrated in Figure 2, an MLP with a single hidden layer includes two matrices, W1 and W2, each of which is linked with numbers that weight inputs from neurons in each layer. The bias components are given by the vectors $-b_1$ and $-b_2$, as you can see.

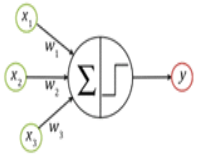


Figure 3.1: Linear threshold unit. It processes the inputs by computing the linear combination of the inputs and applying a step function.

Figure 1 Linear threshold Unit

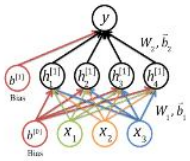


Figure 3.2: Multi-layer perceptron. W_1 and $\vec{b_1}$ denote the weight matrix and the bias vector associated with the input layer. W_2 and $\vec{b_2}$ are the weight matrix and the bias vector associated with the hidden layer.

Figure 2 Multi layer perception..

B. Gradient Descent

Finding the ideal values for the weights may be as if it were an issue of efficiency Its goal is to locate the variables that reduce the error function to the smallest possible value. The error function can be defined as the sum of squared errors is a well-known error function, as given in the equation below, where W, d, and D represent weights, a single training example, and all training instances, respectively. A summation of squared of error between the measured and expected labels (td and yd) is generated for each batch of training data. We must explore the weight space in order to discover the values that minimise this function.

The error function, for example, may be reduced using a standard optimization technique known as gradient descent. Let's pretend we have a sign that quantifies the weight value of its inputs as follows:

Where b, z, and are the bias word, transfer function, target output, and support vectors of something like the universal format, respectively. and the output of The error function of this unit is

The weights are adjusted in the reverse direction as the error probability gradient when utilising the iterative gradient descent approach.

where w is all the proportional control, that defines the refresh stage, and wi is the value associated also with input's ith component. The back - propagation algorithm approach employs optimization algorithms to locate overall minimum distance of a multi-layered convolution (MLP) program's function f(x).. When the step function is substituted for it, σ(x) = 1 1 + ex is used, which can be differentiated.

Convnets (CNNs) are a branch of artificial intelligence that uses

C. CNNs

(convnets) are one of the most widely used deep learning techniques. in areas including object recognition [18], recognize faces [19], handwriting identification [20], audio acknowledgment [21], and natural language processing [22]. Convolution, a mathematical procedure, for example, is used in these platforms. CNNs are formed up of three main building pieces as a rule: a convolutional structure (or convolutional-like) We'll go over these construction blocks, as well as some basic concepts like softmax and dropouts, in the following part..

D. Convolutional Layer

Convolutional layers of CNNs compute the output by convolution rather than multiplication. As a byproduct, the axons in the convolution have increased in number. aren't all linked to the neurons in the layers above them. This design was inspired by a local receptive field discovered on neurons in the visual brain. To put it another way, the neurons are trained to respond to stimuli that are limited to a certain location and structure. The number of parameters in deep neural networks is significantly decreased because of convolution due to the sparse connection and parameter sharing generated as a result of convolution. The convolution of a 2x2 matrix with a 3x3 evocative channel is shown in Figure 3. The output has a volume of 2 2 1. In general, the output size is (nh-f + 1) (nh-f + 1) nf, where nh represents the input height, nw represents the input width, and nf represents the number of kernels. The kernel's depth is determined by the depth of the input. In the case presented in Figure 4, the input depth is nc = 1. As a result, the kernel depth is 1 as a result of this. Although the broadcast depth is one, there is only one kernel. As can be seen, each piece of data in a CNN with weak connections is the sum of something like the hidden neurons in the associated visual field. The foundation is also distributed across the layer, making it possible for CNNs to interchange information. The number of steps made by the kernel as it advances along the input is called stride. In our case, the stride is s = 1, indicating that the kernel moves one step across the picture. It's worth noticing that the input volume decreases with each convolutional layer. To prevent this decrease, we can pad the inputs outside boundary with zero.

E. Pooling Layer

The convolution operation is the second core part of CNNs. This layer is designed to reduce the products' sensitivity to local switching element. In some cases, this inverse to slight local translation might reduce positioning accuracy and lead to misdiagnosis. Pooling can assist CNNs extract the characteristics of interest more efficiently when specific geographical features aren't required. By lowering the number of dimensions and variables, pooling can also help to prevent fitting problem. During pooling, a subsample is collected from the outputs. A kernel (rectangle reception field) is used in both decoding and fully convolutional to gather the data of the synapses within the swimming disk using a clustering algorithm such as maximal, averaged, or enhanced mean. The number of pooling cores, tilting move, and buffering must all be established before a similar technique can be used in CNNs. Highest sharing for a series of blocks using a 2x2 swimming filter that travels one pixel out across column is shown in Figure 4. (ie., stride of 1).

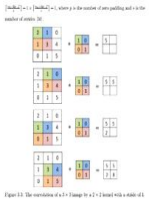


Figure 3 The convolution of 3x3 images by 2x2 kernel with stride of l

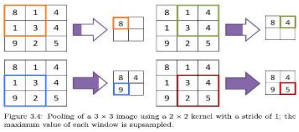


Figure 4 Pooling of 3x3 image using a 2x2 kernel with a stride of l: maximum window is supsampled

F. Fully Connected Layer

A conventional CNN, for example, includes many After each convnet, there is a multilayer perceptron. Finally, CNNs have a totally connected layer, which in its most basic form is a normal MLP. The inputs are now either abstracted or categorised depending on attributes recovered by the preceding layers as a consequence of subsequent processing of the characteristics. Softmax Unit 3.2.4 The outputs of the totally connected layer is usually a svm unit. A convolution layers unit depicts the probabilities of classifiers that used the nonlinear activation function. (Normalised exponential). The softmax function, to be more specific, reflects the probabilistic model of two possible classes. Equation following shows one example of this.

With z is the k-way svm unit's result, softmax()i is the likelihood that perhaps the input occurrence belongs to class I and zi is the vector z's ith member.

Cross Entropy

The cross-entropy between labels supplied by training data and outputs formed by the softmax activation function is used to calculate the loss function. The isolated system is a gauge of its sophistication. a system.

Where l indicates the real probabilistic model of the tags and s signifies the softmax unit's estimated probability distribution. Dkl = divergence kullback entropy H(l) From the left, Leibler s.

Entropy of data in discourse analysis that indicates the with a business, the extent of ambiguity or option. Entropy is a term that may be defined as

When a probability distribution, p, is given to a stochastic process, X, with potential values of x1, x2,.....,xn,

To put it another way, the Kullback-Leibler divergence is a metric for how well a classification predicts. When s deviates from a Cross Entropy (DKL) minimization is the primary aim of learning, it becomes increasingly difficult to predict the future. The discontinuous Thermodynamics can be represented in many different of ways. as

Convolutional neural networks (CNN) assess the probabilistic model of classes using the soft max unit, and bridge is a better error term than mean squared error since mean squared error is more effective in regress scenarios when the output has genuine continuous value.

G. Mini-batch Learning

This is because deep-learning models have many training examples, which slows down the learning process. The use of small batches of samples from the training set seeks to solve the problem of data abyss. To put it differently, instead of training the network with a single iteration's worth of data, the optimization issue involves several error subfunctions rather than a There is just one input parameter.

Learning in small groups is a type of learning that sits between batch and stochastic learning. In batch learning, which we mentioned before For each repetition, all learning algorithms are also used to construct the system. In the probabilistic world, education, on the other hand, each instance is utilised individually to produce an error function that is then used to train the model. The main advantage of stochastic and mini-batch learning approaches is that they save time and computational resources. Stochastic or mini-batch methods, on the other hand, never attain the global optimum; rather, they approach it or fluctuate about it. Whereas a higher amount of the dataset is taken into account while improving the error measures, ’s micro learning is more correct than probabilistic education. On either hand, unpredictable training is more prone than micro batch learning to prevent global optimum. In deep learning models like convolutional layers (CNNs), mini groups are commonly 64, 128, 256, 512, or 1024.

H. Dropout

Color filter education has diametrically opposed character traits: diversity plus bias. Based on the learning set used, the approximated basic theory of a systems can be altered. They're known as variations, which is a term that refers to different ways of doing things. It is desirable to have an overall system model that is less sensitive to the individual training set to accomplish this. Models with large variance, which follow sound changes in training examples, produce overfitting. Models with a large variance, on the other hand, perform On the test dataset, it worked fine, but not on test set. Two typical historical strategies for addressing dimensionality of the data are L1-norm and L2-norm reactivating.

As can be seen in samples as below L1-norm levelling and L2-norm consolidation, the goal apply the results load degradation into in the gradient descent.

l I is just the completely accurate classification of the ith training dataset, s I is the presumed label of something like the ith learning algorithm, w is the key differentiators, and is a hyper factor that determines the homogenisation confidence, where m is the mass of dataset, L is the linear filter, such like cross free energy or mse error (MSE), l I has been the true identifier of the ith base classifier, s I is the estimated identifier of the ith test case, w is the. The optimisation is restricted to small values of w when the regularisation term is included. As a consequence, the model's intricacy may be reduced by assigning weights to a smaller number of features.

V. ARCHITECTURAL SETUP

We used deep learning techniques (CNNs) to characterize uttered phrases content - based in the latest research. We used a private collection to train the classifier our systems in addition to numerous well-known criteria for voice utterance verification. TensorFlow (an open-source framework javascript and C++) served as the programming platform for our CNN models. The experimental procedure involved in the present investigation is described in detail.

A. *Pre-processing*

Prior to any processing, all claims were recomputed and filtered just use an attributing to the fact FIR bandpass filter at a bandwidth rate of 16 kHz to guarantee a constant sampling rate through all collections. The sub - bands of all acoustic words also were created. A frequency spectrum is a visual representation of energy fluctuations at frequency modulation across time. Time is represented by the primary axis (horizontal axis), while rate is represented by the axial (vertical axis). The amount of darkness or the colours are used to convey the energy or power.

Spectrograms are classified into two types:

1) Wide-band spectrograms

2) Narrow-band spectrograms

Wide-band spectrograms have a better temporal resolution than narrow-band spectrograms. Because of this property, individual glottal pulses may be observed in wide-band spectrograms. The sample rate of narrow-band spectrograms is higher than that of wide-band spectrograms. This feature allows single harmonics to be resolved in narrow-band spectrograms.

Figure 5 depicts the wideband and narrowband spectrograms of a voice utterance. Because of relevance of vocalisations and the We chose to convert all utterances into wide-band speech signals due to the fact that its phonetic pulse is associated to one interval of sound. The Bigging window height was set to 5 milliseconds, resulting in a 4.4 millisecond overlap. In all, 512 DFT points were employed in this investigation. Furthermore, spectrograms with harmonics greater than 4 kHz were removed since, in many cases, rates below 4000 Hz are acceptable for auditory processing. In testing, removing energy over 4000 Hz increased the classifiers' productivity. As a consequence, there seem to be a minimum of 129 operating frequencies. All power spectral pictures were downsized to 129 bits plus notation to also have a average of 0 and a standard deviation of around one.

B. *Architecture*

The deep neural network employed in this study had a base architecture ofAs shown in Figure 6, a fcn includes hidden units and one dense layer with 1024 nodes in the input layer. A 5-way or 7-way softmax unit was utilized to assess the prediction algorithm of the courses, limit on the amount of classifications. A minimum or median layer was implemented out after every cnn model. Linear Transfer Units (ReLU) were utilised as kernel functions in multilayer and strongly connected layers to offer non-linearity to the model. For fully connected layers, the starting input size was set to 5x5 with either a phase of 1. In the first and 2nd convolution operation, the numbers of kernels got set at 8 and 16, respectfully. The dipping layers' nucleus size was set to 2x2 with a stride of 2. Over the education data's micro groups, the Particle swarm optimization was designed to decrease the loss function, which was trans. The tiny bunches were set to 512 bytes in size. A total of 100 training sessions were held. Using Visuals Units, the models in this existing situation anywhere from 30 minutes to two days to train (GPUs). GPUs, then instead of Computers, are utilised to speed up computation since they contain many cores and can operate a high number of repeated threading. The Crane cluster somewhere at University of Nebraska-Holland Lincoln's Information Center was used for our study. The ejection strategy was also implemented into the fully linked layer to increase network efficiency when appropriate..

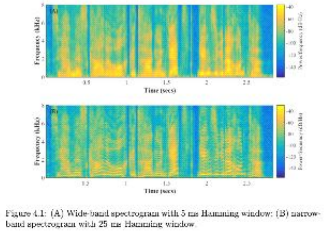


Figure 5(A) Wide band spectrogram with 5ms hamming window (B) Narrow band spectrogram with 25ms hamming window

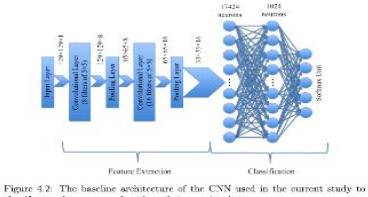


Figure 6The baseline architecture of the CNN used in the current study to classify speech utterance based on their emotional states

C. Sets for Training and Testing

5-fold trans was used to train and evaluate our models. To put it another way, the knowledge was subdivided into four folds. Our networks were validated on the first bend, while the others were utilised as a test set. The middle fold was utilised to test our predictions, while an subsequent folds were also used for education and other uses. The test data were enhanced by introducing white Gaussian noise with both a +15 signal (SNR) to each speech signals either six times or 20 times to minimise prediction error and the undesirable effect of low collection volumes. To test our methods, I used the actual data, which became noise-free. The greater the amount of data available, the more it was used for training. Following that, one-hot gradients were used to represent the labels of the training and test sets. The total quantity iterations ranged from 100 to 4000. Due to compute and strict curfews, the efficient training period was chosen at 100.

VI. RESULTS AND OBSERVATIONS

The following steps have been undertaken to obtain the required results:

Firstly the required libraries were imported to get access from python to any other module.as shown in Figure 7

```
In [1]: import os
        import librosa
        import IPython.display as ipd
        import numpy as np
        from scipy.io import wavfile
```

Figure 7 importing the required libraries

Then sample recording for maximum rate (16000) was checked as demonstrated in Figure 8 and then the rate was changed to half (8000) as shown in Figure 9.

```
In [2]: trainDir = '/Users/tahirshowkatbazaz/Desktop/Speech CNN/train/audio/'
        samples, sample_rate = librosa.load(trainDir + 'cat/0ab3b47d_nohash_0.wav', sr = 16000)
        ipd.Audio(samples, rate=sample_rate)
        print(sample_rate)

        16000
```

Figure 8 Checking the sample rate of recordings

```
In [3]: samples = librosa.resample(samples, sample_rate, 8000)
        ipd.Audio(samples, rate=8000)
Out[3]: ▶ 0:00 / 0:01
```

Figure 9 Changing the sample rate to 8000 and checking its audibility

Figure 10 shows the recording labels; after that, all files with the wav suffix were filtered and their sample rate was raised to 8000. Then, as shown in Figure, all wave recordings were appended to the wave list and all labels to the all label list.

```
In [4]: labels=["bed", "bird", "cat", "dog", "down", "eight", "five", "four"]
```

Figure 10 Labels of the recordings

```
In [*]: trainDir = '/Users/tahirshowkatbazaz/Desktop/Speech CNN/train/audio/'
        all_wave = []
        all_label = []
        for label in labels:
            print(label)
            waves = [f for f in os.listdir(trainDir + '/'+ label) if f.endswith('.wav')]
            for wav in waves:
                samples, sample_rate = librosa.load(trainDir + '/' + label + '/' + wav, sr = 16000)
                samples = librosa.resample(samples, sample_rate, 8000)
                if(len(samples)== 8000) :
                    all_wave.append(samples)
                    all_label.append(label)
```

Figure 11 Appending recordings in all wave list and all labels in all label list

Following that, all labels and recordings were coded in machine language, and the data was divided into two groups, as shown in Figure 12.

Figure 12 Encoding all labels and recordings and split data to train and test

The ID CNN Model was then created. As demonstrated in Figure, Max pooling and drop out layers were added as needed. 13 The output was then sequenced using Flatten, and an output model was created, as illustrated in Figure 14.

Figure 13 Adding of max pooling and dropout layers

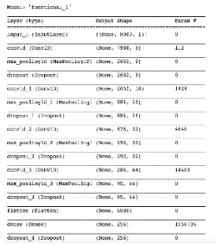


Figure 14 Sequenced output

The model was then trained, and after many epochs and the least amount of validation accuracy, a model was stored, as illustrated in Figure 15.

Figure 15 Training the model

Finally, testing is carried out by uploading the same model and testing a random record from the data; the results of the testing revealed that 95% of the data was right.

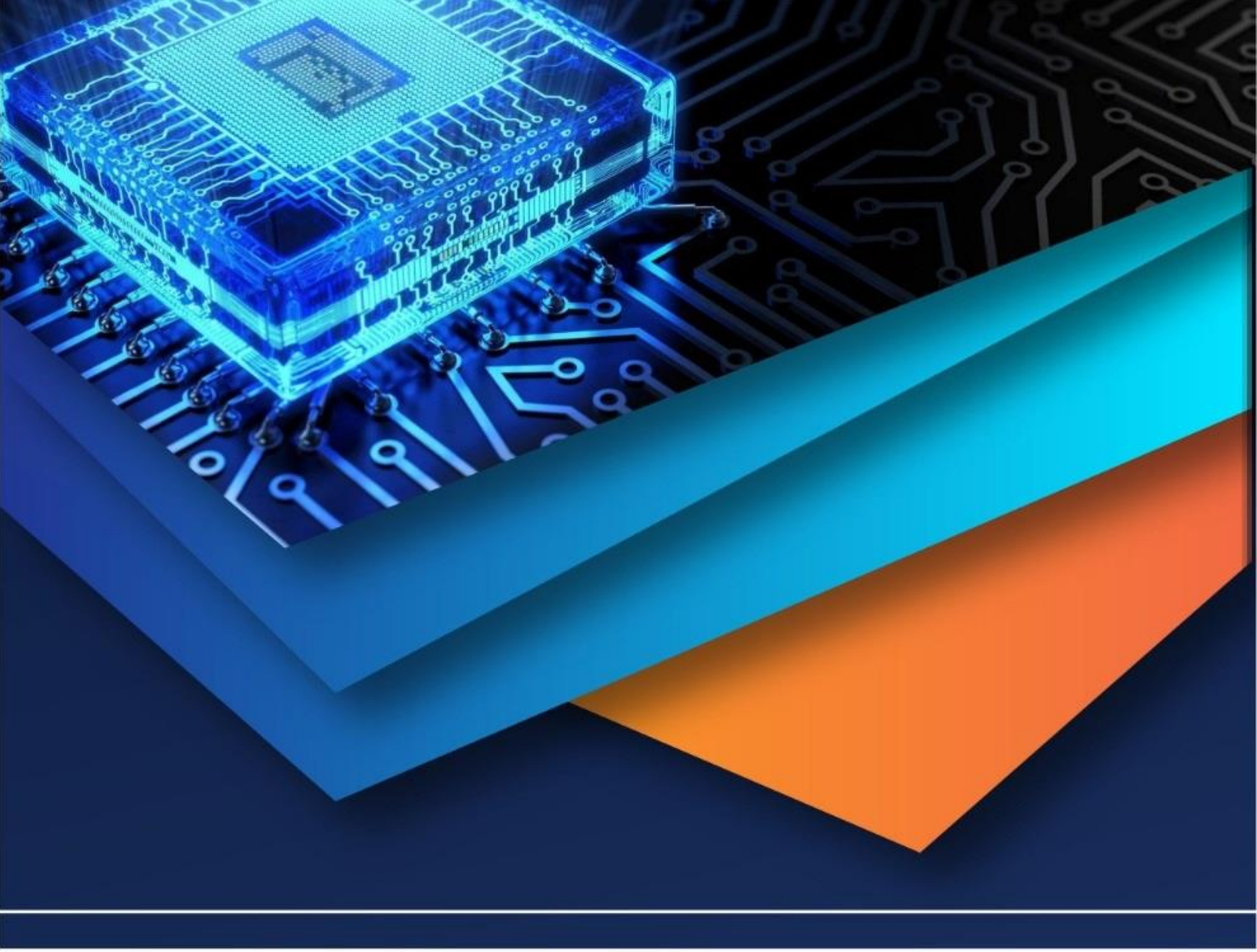
Figure 16 Testing the model

VII. CONCLUSION

The level of security provided by a voice recognition is its primary function and value. The way people do business on the internet is changing thanks to speech recognition. Organizations can create and implement voice-enabled online products right now since voice mail technology combines speech recognition with telephony. ANN is used to recognise speech in this study. Artificial neural networks (ANNs) are thought to be one of the most promising technologies. If you're looking to categorise speech, this is a good place to start. Conventional logic doesn't work as well with them as it does with other people. Because of its low sensitivity, we chose MLP Architecture, which has shown to be the ideal option. Scientists are increasingly interested in learning more about voice recognition, which has had a tremendous influence on society. This technology has a bright future ahead of it, and hardware development is critical to its success.

REFERENCES

[1] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. Artificial intelligence: a modern approach, volume 2. Prentice hall Upper Saddle River, 2003.
[2] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. Neuron, 95 (2):245–258, 2017.
[3] Marcel Van Gerven. Computational foundations of natural intelligence. Frontiers in Computational Neuroscience, 11:112, 2017.
[4] Paul R Cohen and Edward A Feigenbaum. The handbook of artificial intelligence, volume 3. Butterworth-Heinemann, 2014.
[5] Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010.
[6] Marvin Minsky. The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind. Simon and Schuster, 2007.
[7] Dong Yu and Li Deng. AUTOMATIC SPEECH RECOGNITION. Springer, 2016.
[8] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.
[9] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. A maximum likelihood approach to continuous speech recognition. In Readings in speech recognition, pages 308–319. Elsevier, 1990.
[10] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. The Bell System Technical Journal, 62(4):1035–1074, 1983.
[11] Su-Lin Wu, ED Kingsbury, Nelson Morgan, and Steven Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, volume 2, pages 721–724. IEEE, 1998.
[12] Vaibhava Goel and William J Byrne. Minimum bayes-risk automatic speech recognition. Computer Speech & Language, 14(2):115–135, 2000. 71
[13] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing, 13(2): 293–303, 2005.
[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
[15] Aur´elien G´eron. Hands on machine learning with scikit-learn and tensorflow, 2017. 73
[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[17] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks, 8(1):98–113, 1997.
[18] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems, pages 396–404, 1990.
[19] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. The microsoft 2016 conversational speech recognition system. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 5255–5259. IEEE, 2017.
[20] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Advances in neural information processing systems, pages 649–657, 2015.
[21] David H Hubel. Single unit activity in striate cortex of unrestrained cats. The Journal of physiology, 147(2):226–238, 1959.
[22] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology, 148(3):574–591, 1959.
[23] Pierre Baldi and Peter J Sadowski. Understanding dropout. In Advances in neural information processing systems, pages 2814–2822, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY